Understanding
Distributions of Performance

Dr. Kenneth W. Regan, Univ. at Buffalo
Bartlomiej Macieja, Warsaw, Poland
Guy Haworth, Univ. of Reading (UK)

ACG13, Tilburg, The Netherlands, Nov. 2011
Back story

- A sequence of papers on ‘Assessing Chess Players’
  - (Deeper) Model Endgame Analysis (2003, 2005)
  - Reference Fallible Players (2007)
  - Skill Rating by Bayesian Inference (2009) ... IEEE CIDM ‘09
  - Performance and Prediction, (2009) ... ACG12, Pamplona
  - Intrinsic Chess Ratings (2011) ... AAAI-11, San Francisco

- Topics
  - The creation of a Skilloscope to rank players
  - Comparison of and correlation with ELO scales
  - Detection of plagiarism ... and ELO Scale instability
The focus today

- the question of *ELO Inflation*
- common views about the FIDE ELO scale
  - View 1: *ELO 2700 means lower quality play today*
  - View 2: *ELO 2700 should mean ‘best few’ players*
  - it is impossible for ELO to conform to both views over time

- Three-dimensional assessment of the inflation question
  - Population dynamics
  - ‘Average Error’ in categorised FIDE tournaments
  - Parametric models $A(s, c)$ of Virtual ELO players
    Use of these $A(s, c)$ to assess tournament (players) etc
Summary Results

- **Population Analysis**
  - the figures do not provide evidence of inflation
  - nor do they disprove the ‘inflation theory’ but ...
    - they do exclude two sources of inflation

- ‘Average Error’ calculations on FIDE-rate tournaments
  - Single-PV analysis picks out ELO-levels of competence
  - show some signs of deflation in the last 20 years
    - i.e. improving standards at ELO Level ‘E’ (for high ‘E’)

- Modelling players using statistical regression:
  - Multi-PV analysis acknowledging most relevant options
  - the ‘optimal parameters’ are reasonably stable over time
1. Population dynamics

Malthus (1798): Darwin, Wallace
Verhulst (1838)
Population analysis

- What factors account for the increase in ELO 2203+ players?
  - Inflation or other factors
  - Verhulst (1838): $\frac{dP}{dt} \propto P.(N–P) \propto P.N – P^2 = a.P – b.P^2$
  - This is the Logistic Curve

- the actual data fits well to a Logistic Curve

- The ‘fit’ supports the idea that:
  - standard population theory explains ELO-population growth
  - the ELO population is not shifting up the scale
  - The ELO population is not expanding up the scale

- ... no support for ELO Inflation Theory
Players above ELO 2203 v Logistic Curve

Figure 1: Growth of number of players rated at least 2203 since 1971
2. Single PV Analysis of Player ‘error’

- RYBKA 3.0 1-cpu run in single-PV mode to 13-ply depth
  - Larry Kaufman estimated depth 14 = 2750
  - We estimate our engine at 2650-2700 (2900 ... 2400)
- Run manually in Arena GUI (versions 1.99, 2.01).
  - reproducible except when Rybka ‘stalls’
- All tournaments of category \( \geq 11 \) analysed
  - moves 1-8 ignored; positions > ‘3.00’ ignored
  - 3.77m of 4.00m+ moves analysed
  - two 4-core PCs employed ...
- The data is quorate and results seem robust
- Large-scale data needed as benchmark in ‘anti-cheating’ cases
Average Error

- When played move $\neq$ Rybka’s first move, 
  \[ \text{error} = \max(\text{value} - \text{value(next position)}, 0). \]
- This is logistically simple: perhaps better to use \text{value(next at depth 12)}
- Details differ from Guid/Bratko’s work
  \[ \text{hence we label our errors ‘AE’ rather than their ‘AD’} \]

- A comparison of Average Error against Position Value
  \[ \text{larger errors are made in more decisive positions} \]
  \[ \text{suggests a scaling … } 1/(1 + |\text{position value}|) \]
Average Error v Position Value

E5 = 1970—1984
E6 = 1985—1999
E7 = 2000—2009
E8 = 2010—2019
Average Error by Move Number

The effect of time pressure approaching move 40 is clear.

Moves 17—32 bridge between opening theory and the worst of Zeitnot
Plot of Scaled Average Error by Category

Plot lines would slope up if there were considerable rating inflation.

Some evidence of deflation in higher categories.

Curves are similar to case of ‘all moves’; error itself is a little higher.

Overall no-inflation verdict thus independent of today’s greater opening theory knowledge.
3: Parameterised Models of Players

- **Motivation**
  - *Average Error* does not use the decision’s full context
  - Predicting a player’s move requires an ‘agent’ model of the *fallible player* at their skill level
  - Hence the need for a range of *Reference Fallible Players*

- $\mathbf{A}(\mathbf{c})$ is an agent with *behaviour parameters* $(c_1, c_2, \ldots)$
  - current model has two parameters:
    - $s \equiv \text{sensitivity}, c \equiv \text{competence}$
  - $\text{Prob } [ \mathbf{A}(\mathbf{c}) \text{ chooses move } m_i ] \propto PF(\text{posval}, v_i, c)$
  - $s$ and $c$ determined by statistical regression
‘Statistically fitting’ agents to human players

- Population used here are ‘Virtual ELO Players’, ELO $E$
- $E = 2700, 2600, 2500$ etc
- Virtual players are composite of actual players who ...
  - Are within 10 ELO of, e.g. 2400 and playing a ‘like’ player

- $m_0$ is the move with the best computer evaluation $v_0$
- $m_i$ is the $i$th best move and has value $v_i \leq v_{i-1}$
- $\delta_i$ is a scaling of $v_i - v_0$
- the probability function $PF$ is defined by:
  \[
  \log(p_i)/\log(p_0) = e^{c(-\delta/s)^c}
  \]
  this function seems likely to be the best of those defined
Results from defining agents $A(s, c)$

- for the Virtual ELO 2400 player, e.g., we define $A(s, c)$
- $A(s, c)$ also has an Average Error $AE_c$
- thus, we may associate $Ae_c$ with ELO 2400

- now, given a set of players in a tournament...
- We may determine an $A(s, c)$ for the tournament
  ... And indeed, for each player in the tournament
- Thus, we may determine a ‘performance ELO’
  for the tournament and each player
- These may be compared with the average FIDE ELO
  for the tournament, and the TPR for each player
A(s, c) results on the training sets

<table>
<thead>
<tr>
<th>Elo</th>
<th>s</th>
<th>c</th>
<th>IPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2700±10</td>
<td>.078</td>
<td>.503</td>
<td>2690</td>
</tr>
<tr>
<td>2600±10</td>
<td>.092</td>
<td>.523</td>
<td>2611</td>
</tr>
<tr>
<td>2500±10</td>
<td>.092</td>
<td>.491</td>
<td>2510</td>
</tr>
<tr>
<td>2400±10</td>
<td>.098</td>
<td>.483</td>
<td>2422</td>
</tr>
<tr>
<td>2300±10</td>
<td>.108</td>
<td>.475</td>
<td>2293</td>
</tr>
<tr>
<td>2200±10</td>
<td>.123</td>
<td>.490</td>
<td>2213</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Elo</th>
<th>s</th>
<th>c</th>
<th>IPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2700±10</td>
<td>.079</td>
<td>.487</td>
<td>2630</td>
</tr>
<tr>
<td>2600±10</td>
<td>.092</td>
<td>.533</td>
<td>2639</td>
</tr>
<tr>
<td>2500±10</td>
<td>.098</td>
<td>.500</td>
<td>2482</td>
</tr>
<tr>
<td>2400±10</td>
<td>.101</td>
<td>.484</td>
<td>2396</td>
</tr>
<tr>
<td>2300±10</td>
<td>.116</td>
<td>.480</td>
<td>2237</td>
</tr>
<tr>
<td>2200±10</td>
<td>.122</td>
<td>.477</td>
<td>2169</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Elo</th>
<th>s</th>
<th>c</th>
<th>IPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2600±10</td>
<td>.094</td>
<td>.543</td>
<td>2647</td>
</tr>
<tr>
<td>2500±10</td>
<td>.094</td>
<td>.512</td>
<td>2559</td>
</tr>
<tr>
<td>2400±10</td>
<td>.099</td>
<td>.479</td>
<td>2397</td>
</tr>
<tr>
<td>2300±10</td>
<td>.121</td>
<td>.502</td>
<td>2277</td>
</tr>
</tbody>
</table>

Inflation would show as IPR > Elo in tables at right. Pretty much none.
### Some recent tournaments

<table>
<thead>
<tr>
<th>Event</th>
<th>cat</th>
<th>Elo</th>
<th>IPR</th>
<th>Diff</th>
<th>Event</th>
<th>cat</th>
<th>Elo</th>
<th>IPR</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linares 1993</td>
<td>18</td>
<td>2676</td>
<td>2522</td>
<td>-154</td>
<td>Corus 2007</td>
<td>19</td>
<td>2717</td>
<td>2763</td>
<td>+46</td>
</tr>
<tr>
<td>Linares 1994</td>
<td>18</td>
<td>2685</td>
<td>2517</td>
<td>-168</td>
<td>Mexico 2007</td>
<td>21</td>
<td>2751</td>
<td>2708</td>
<td>-43</td>
</tr>
<tr>
<td>Dortmund 1995</td>
<td>17</td>
<td>2657</td>
<td>2680</td>
<td>+23</td>
<td>Sofia 2007</td>
<td>19</td>
<td>2725</td>
<td>2576</td>
<td>-149</td>
</tr>
<tr>
<td>Dortmund 1996</td>
<td>18</td>
<td>2676</td>
<td>2593</td>
<td>-83</td>
<td>Sofia 2008</td>
<td>20</td>
<td>2737</td>
<td>2690</td>
<td>-47</td>
</tr>
<tr>
<td>Dortmund 1997</td>
<td>18</td>
<td>2699</td>
<td>2639</td>
<td>-60</td>
<td>Sofia 2009</td>
<td>21</td>
<td>2754</td>
<td>2703</td>
<td>-51</td>
</tr>
<tr>
<td>Dortmund 1998</td>
<td>18</td>
<td>2699</td>
<td>2655</td>
<td>-44</td>
<td>Nanjing 2010</td>
<td>21</td>
<td>2766</td>
<td>2748</td>
<td>-18</td>
</tr>
<tr>
<td>Dortmund 1999</td>
<td>19</td>
<td>2705</td>
<td>2749</td>
<td>+44</td>
<td>Shanghai 2010</td>
<td>21</td>
<td>2759</td>
<td>2829</td>
<td>+70</td>
</tr>
<tr>
<td>Sarajevo 1999</td>
<td>19</td>
<td>2703</td>
<td>2722</td>
<td>+19</td>
<td>Bilbao 2010</td>
<td>22</td>
<td>2789</td>
<td>2904</td>
<td>+115</td>
</tr>
<tr>
<td>San Luis 2005</td>
<td>20</td>
<td>2738</td>
<td>2657</td>
<td>-81</td>
<td>Moscow 2010</td>
<td>21</td>
<td>2757</td>
<td>2690</td>
<td>-67</td>
</tr>
<tr>
<td>Corus 2006</td>
<td>19</td>
<td>2715</td>
<td>2736</td>
<td>+21</td>
<td>London 2010</td>
<td>20</td>
<td>2725</td>
<td>2668</td>
<td>-57</td>
</tr>
<tr>
<td>Sofia 2006</td>
<td>20</td>
<td>2744</td>
<td>2744</td>
<td>0</td>
<td>Averages</td>
<td>19</td>
<td>2722</td>
<td>2690</td>
<td>-32.6</td>
</tr>
</tbody>
</table>

**IPRs are reasonable; half of shortfall is from Linares 1993-94.**

**No support for inflation hypothesis here either.**
The Canadian Open, July 9-17, 2011

- 9 round Swiss: 149 players (115 with FIDE ratings)
- 623 games available and analysed (of 647 played)

<table>
<thead>
<tr>
<th>Whole event</th>
<th>CanR</th>
<th>TPR</th>
<th>IPR</th>
<th>Restrict</th>
<th>CanR</th>
<th>FIDE</th>
<th>IPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>2144</td>
<td>2142</td>
<td>2117</td>
<td>to 115</td>
<td>2211</td>
<td>2139</td>
<td>2203</td>
</tr>
<tr>
<td>St. Deviation</td>
<td>258</td>
<td>261</td>
<td>379</td>
<td>FIDE-</td>
<td>229</td>
<td>220</td>
<td>345</td>
</tr>
<tr>
<td>Wtd. by games</td>
<td>2156</td>
<td>2154</td>
<td>2134</td>
<td>rated</td>
<td>2221</td>
<td>2147</td>
<td>2219</td>
</tr>
<tr>
<td>Wtd. by moves</td>
<td>2173</td>
<td>2172</td>
<td>2161</td>
<td>players:</td>
<td>2236</td>
<td>2161</td>
<td>2242</td>
</tr>
</tbody>
</table>

- Can compare IPRs with TPRs and with FIDE ELO ratings
- Impression is that Canadian players here are too low in FIDE ELO
Conclusions

- Three-dimensional assessment of the *ELO Inflation* issue
- Population analysis does not support *inflation theory*
- Average Error hints at *deflation* rather than *Inflation*
- Multi-PV analysis is effective on a smaller scale
  - yields credible *Intrinsic Performance Ratings*
  - these IPRs correlate well with ELO
  - ... a vote of confidence for both
The Way Ahead ... Some thoughts

- Improved infrastructure for our computation experiments
  - *repeatability* requires engines which do not *stall*
  - a database to store engine-evaluations of positions
  - automated exploitation of distributed computing resources

- Integration of two statistical approaches
  - Statistical regression
  - Bayesian inference

- Further exploitation
  - Our analyses can be cloud-sourced in real-time
  - Application to other *Fallible Decision Maker* areas
Our thanks to ...

- The ARENA GUI programmers for full-analysis scripting

- Programmers associated with TOGA II and RYBKA

- Univ. Of Buffalo CSE and Univ. of Montreal for support

- Tamal Biswas, managing the data, creating the graphs

- Hugh Brodie and David Cohen for the Can. Open games